CUNY SPS DATA 621 - CTG5 - Final

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

May 23rd, 2019

Contents

1 PROJECT DESCRIPTION AND BACKGROUND

1.1 Background

[Jeremy's writing this up]

1.2 Hypothesis

1.3 Our approach, setup, and workflow

[We should discuss this]

2 DATA PREPARATION

2.1 Cross-validation

[Assuming this requires a little explanation]

2.2 Bootstrap surrogata data

[Jeremy's writing this up]

Per the synthpop packace explanation (https://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf): "The basic idea of synthetic data is tovreplace some or all of the observed values by sampling from appropriate probability distributions so that the essential statistical features of the original data are preserved. The approach has been developed along similar lines to recent practical experience with multiple imputation methods although synthesis is not the same as imputation. Imputation replaces data which are missing with modelled values and adjusts the inference for the additional uncertainty due to this process. For synthesis, in the circumstances when some data are missing two approaches are possible, one being to impute missing values prior to synthesis and the other to synthesise the observed patterns of missing data without estimating the missing values. In both cases all data to be synthesised are treated as known and they are used to create the synthetic data which are then used for inference. The data collection agency generates multiple synthetic data sets and inferences are obtained by combining the results of models fitted to each of them. The formulae for the variance of estimates from synthetic data are different from those used for imputed data."

"Our aim in writing the synthpop package (Nowok, Raab, Snoke, and Dibben 2016) for R (R Core Team 2016) is a more modest one of providing test data for users of confidential datasets. Note that currently all values of variables chosen for synthesis are replaced but this will be relaxed in future versions of the package. These test data should resemble the actual data as closely as possible, but would never be used in any final analyses. The users carry out exploratory analyses and test models on the synthetic data, but they, or perhaps staff of the data collection agencies, would use the code developed on the synthetic data to run their final analyses on the original data. This approach recognises the limitations of synthetic data produced by these methods."

2.3 Synthesis diagnostics

The original Cleveland dataset contains n = 303 observations over ...

3 BUILDING MODELS

Our literature review revealed that of the many approaches that have been taken, certain types of models stand out in terms of their performance.

[Include grid with examples of previous work]

[See notes in literature reviewed and Kaggle projects for suggestions on variable selection and feature engineering for models (links TBC)]

[See also notes on setup and packages for decision tree, random forest, SVM, and Naive Bayes (links TBC]

3.1 Logistic regression

3.2 Decision tree (CHAID or C&RT?)

3.3 Random forest

3.4 Support Vector Machines

3.5 Naive Bayes

4 MODEL REVIEW AND SELECTION

4.1 Comparison of performance between models

[sensitivity, specificity, accuracy, others metrics?]

[Between different techniques and between original dataset and synthesize dataset for each technique]

4.2 Comparison of performance viz. other studies

5 CONCLUSIONS

Table 1: Summary statistics for numerical variables in the original dataset

	n	min	mean	median	max	sd
age	303	29	54.366337	55.0	77.0	9.082101
trestbps	303	94	131.623762	130.0	200.0	17.538143
chol	303	126	246.264026	240.0	564.0	51.830751
thalach	303	71	149.646865	153.0	202.0	22.905161
oldpeak	303	0	1.039604	0.8	6.2	1.161075

Table 2: Summary statistics for categorical variables in the original dataset

	cp	ca	restecg	slope	thal
	0:143	0:175	0:147	0: 21	0: 2
	1: 50	1: 65	1:152	1:140	1: 18
	2: 87	2: 38	2: 4	2:142	2:166
	3: 23	3: 20	NA	NA	3:117
	NA	4: 5	NA	NA	NA

Table 3: Summary statistics for binary categorical variables in the original dataset

	exang	fbs	sex	target
	0:204	0:258	0: 96	0:138
	1: 99	1: 45	1:207	1:165

Table 4: Summary statistics for numerical variables in the synthesised dataset

	n	min	mean	median	max	sd
age	60600	29	54.32906	55.0	77.0	9.063818
trestbps	60600	94	131.59893	130.0	200.0	17.460541
chol	60600	126	246.90715	242.0	564.0	52.406720
thalach	60600	71	149.60807	153.0	202.0	23.219737
oldpeak	60600	0	1.03686	0.8	6.2	1.146516

6 APPENDIX

6.1 Supplemental tables and figures

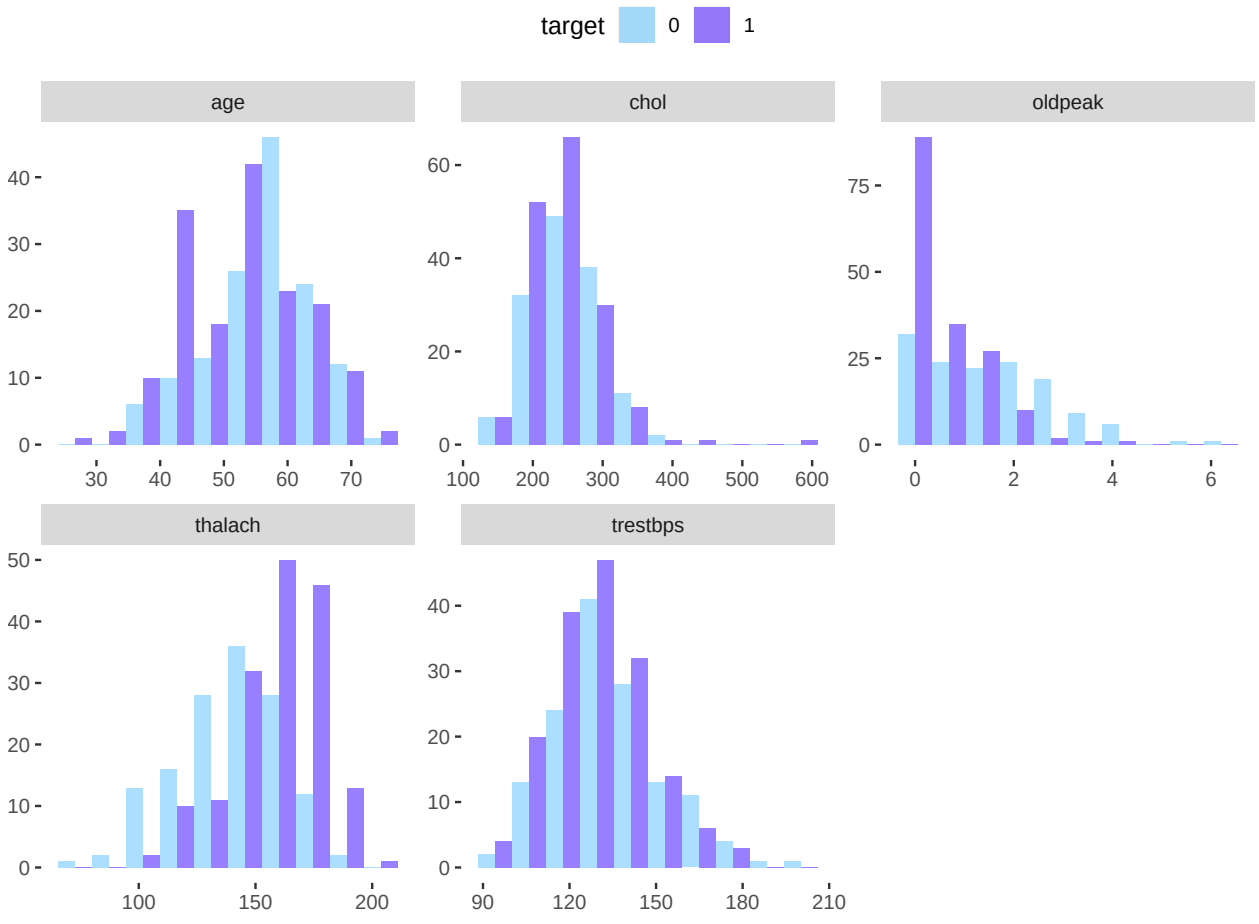


Figure 1: Numeric Data Distributions as a Function of TARGET

Table 5: Summary statistics for categorical variables in the synthesised dataset

	cp	ca	restecg	slope	thal
	0:28570	0:34658	0:30174	0: 4984	0: 369
	1:10135	1:13438	1:29664	1:28183	1: 4029
	2:17207	2: 7515	2: 762	2:27433	2:33157
	3: 4688	3: 3903	NA	NA	3:23045
	NA	4: 1086	NA	NA	NA

Table 6: Summary statistics for binary categorical variables in the synthesised dataset

	exang	fbs	sex	target
	0:40938	0:51613	0:19040	0:29203
	1:19662	1: 8987	1:41560	1:31397

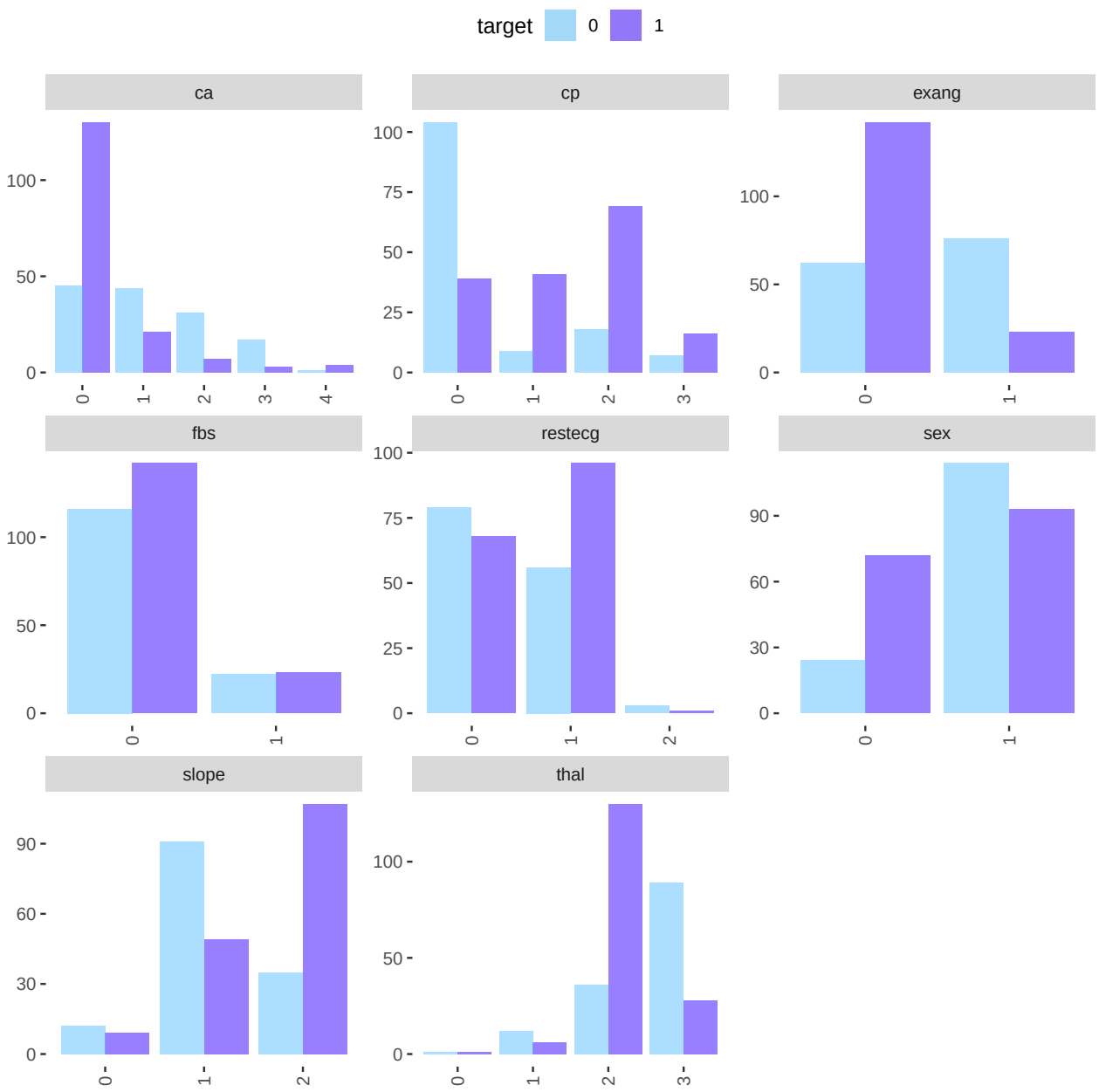


Figure 2: Categorical Data Distributions as a Function of TARGET

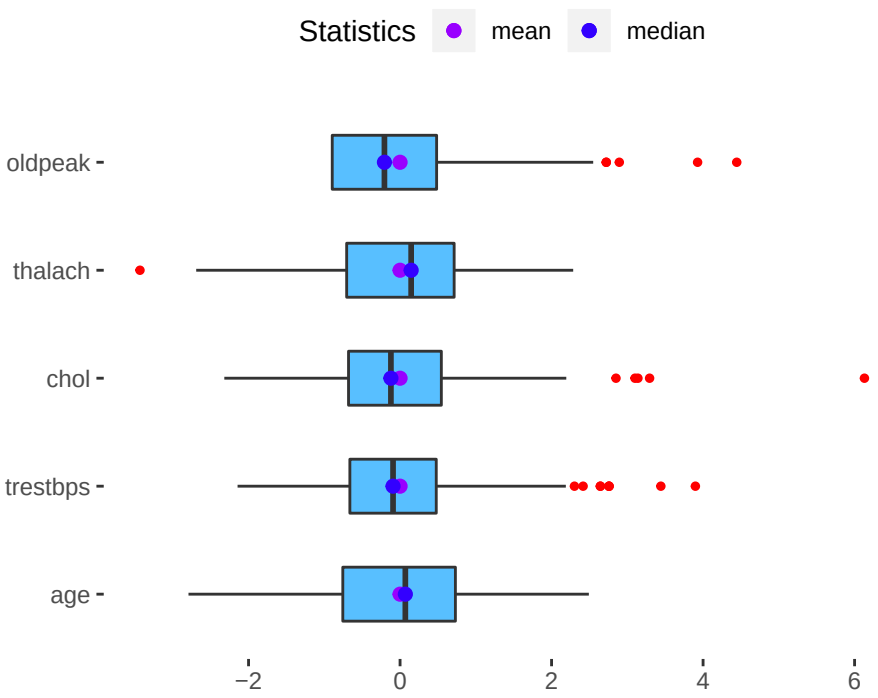


Figure 3: Scaled Boxplots

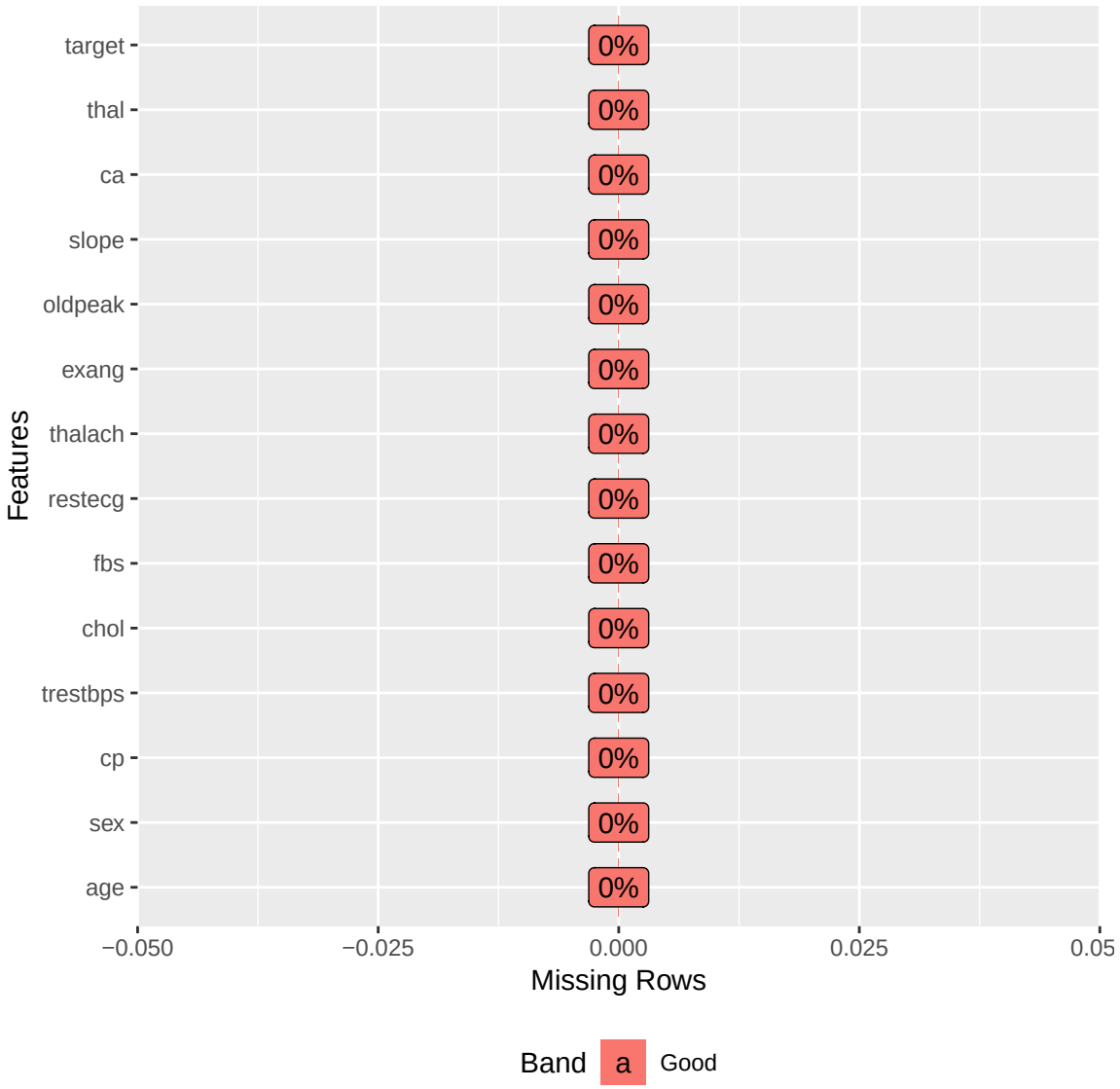
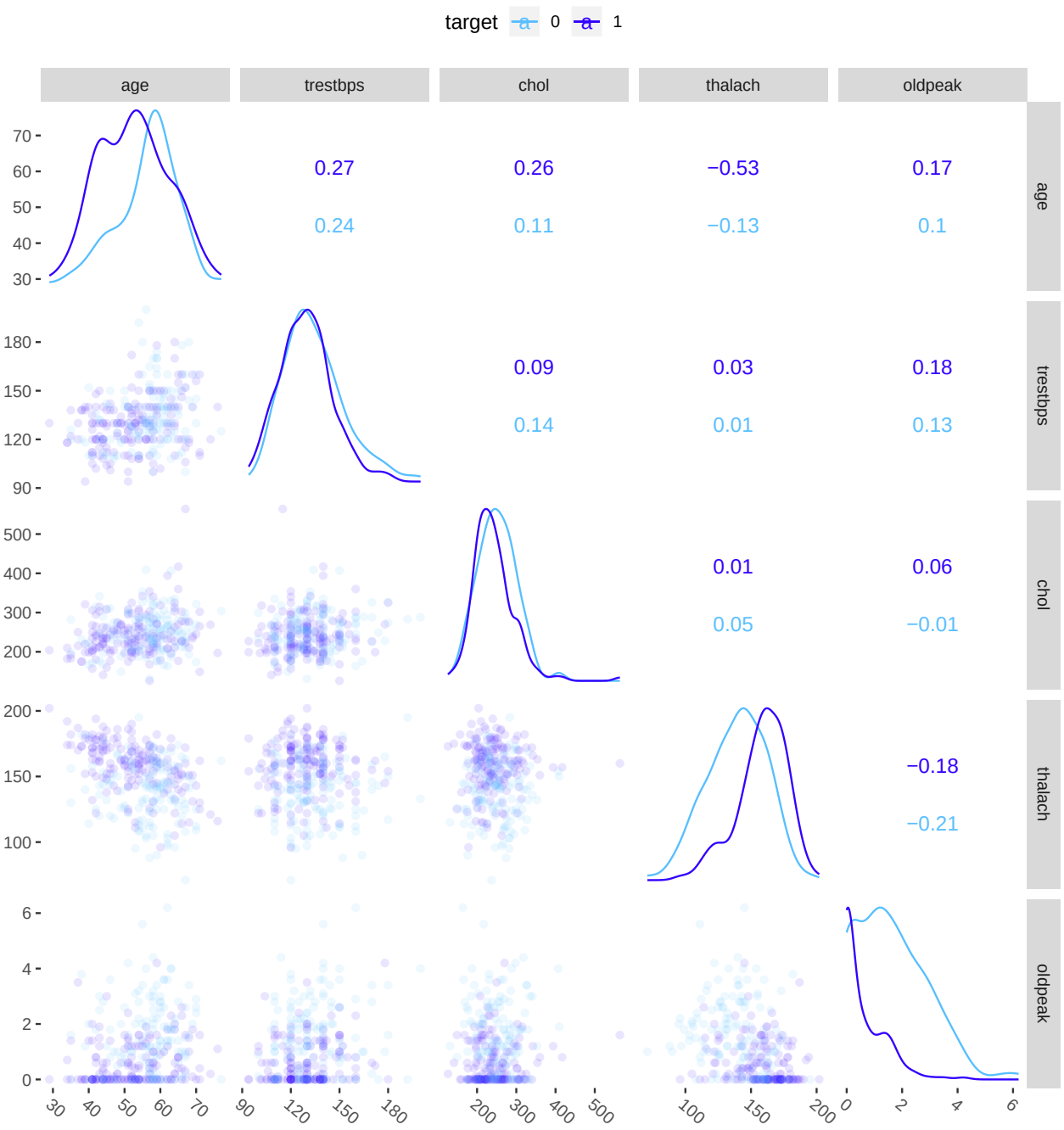


Figure 4: Missing data in the synthesised dataset



6.2 R statistical programming code

The appendix is available as script.R file in `projectFinal_heart` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining